J Mol Med (1997) 75: 694–698

Bioinformatics: Bits and Bytes

by Peer Bork

Pieces of the puzzle: expressed sequence tags and the catalog of human genes

Gregory D. Schuler

The Puzzle of Life

Imagine trying to solve a jigsaw puzzle without having all of the pieces. This is exactly the dilemma faced by researchers in the field of molecular medicine when attempting to understand how human genes and their protein products interact with one another to lead to normal biological functions, how these functions can break down in various disease states, and how normal functions can be restored through molecular intervention. This description of the Puzzle of Life is not meant to deny the importance of environmental and other epigenetic factors, but is simply meant to define the boundaries of a puzzle whose solution is easily within our grasp. To further our basic understanding of human biology and the genetics of inherited diseases, it would be immensely valuable to compile a complete catalog of human gene sequences and to make this information available over the Internet to scientists around the world. Over the past few years huge amounts of data relevant to this puzzle have become available, but solving the puzzle remains a bioinformatics challenge.

Before setting out to solve the Puzzle of Life, it would be useful to have a rough sense of how many pieces it contains. In other words, how many human genes are there? Based on indirect evidence, estimates ranging from approximately 64,000 [1] to 80,000 [2] genes have been advanced. Complete genomic sequencing has been used to generate gene catalogs for several organisms with relatively small genomes [3]. However, sequencing the human genome is a much more daunting task due to its immense size (about 3 billion bases). The United States Genome Project began in 1990 with the ambitious goal of sequencing the human genome within 15 years (i.e., by the year 2005) [4]. Unfortunately, only about 2% of the total bases make up the protein-coding portions of our genes; the remaining 98% is of unknown function and often referred to as “junk DNA.” Thus, sequencing the genome may not be the most efficient way to generate a catalog of human genes. A number of investigators have advocated large-scale sequencing of the transcription products of genes, in the form of complimentary DNA (cDNA) clones, as a prelude to sequencing of the entire human genome. As Brenner [5] put it, “If something like 98% of the genome is junk, then the best strategy would be to find the important 2%, and sequence it first.”

G.D. Schuler
National Center
for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda MD 20894, USA

Please send articles to:
Peer Bork
Max-Delbrück-Center
for Molecular Medicine (MDC)
Robert-Rössle-Strasse 10
D-13122 Berlin, Germany
and:
EMBL
Meyerhofstrasse 1
D-69117 Heidelberg, Germany
E-mail: bork@embl-heidelberg.de
http://www.embl-heidelberg.de/~bork/

An abundance of puzzle pieces

The era of high-throughput cDNA sequencing was initiated in 1991 by a landmark study from Venter and colleagues [6]. The basic strategy involves selecting cDNA clones at random and performing a single automated sequencing read from one or both ends of their inserts. They introduced the term “expressed sequence tag” (EST) to refer to this new class of sequence, which is characterized by being short (typically about 400 bases) and relatively inaccurate (around 2% error). The use of single-pass sequencing was an important aspect of making the approach cost effective. In most cases there is no initial attempt to identify or characterize the clones. Instead, they are identified using only the small bit of sequence data obtained – comparing it to the sequences of known genes and other ESTs. It is fully expected that many clones will be redundant with others already sampled, and that a smaller number will represent various sorts of contaminants or cloning artifacts. There is no point in incurring the expense of high-quality sequencing until later in the process when clones can be validated and a nonredundant set selected.

Despite their fragmentary and inaccurate nature ESTs represent an extraordinarily valuable resource which has accelerated the discovery and study of huge numbers of new genes – new pieces of the Puzzle of Life [7, 8]. In 1992 a database called dbEST [9] was established to serve as a collection point for ESTs, which are then distrib-

uted to the scientific community as the EST division of GenBank [10]. One avenue to gene discovery is to use a database search tool, such as BLAST [11], to perform a sequence similarity search against dbEST. The query for such a search would be a gene or protein sequence, perhaps from a model organism, that is expected to be related to the human gene of interest. Because clone identifiers are carried with the sequence tags, it is possible to obtain the original material to generate a more accurate sequence or to use as an experimental reagent.

Following the initial demonstration of the utility and cost effectiveness of the EST approach many similar projects have been initiated, resulting in an ever-increasing numbers of human ESTs [12–15]. In addition, large-scale EST projects have been initiated in several other model organisms (see Table 1), including *Mus musculus* (L. Hillier et al., in preparation), *Ceanorhabditis* elegans [16, 17], and *Arabidopsis* thaliana [18]. Collectively these efforts have caused the number of EST sequence entries in GenBank to soar past the one million mark in the early part of 1997. Currently 71% of all GenBank entries, 40% of the individual bases, are derived from ESTs. Two projects are worthy of particular mention. The first is a project funded by Merck & Co. and carried out at the Washington University Genome Sequencing Center [15], which has (as of 1 June 1997) contributed 508,945 human EST sequences (69% of all human ESTs in dbEST). The second is a similar effort for mouse EST sequences funded by the Howard Hughes Medical Institute and, again, carried out at the Washington University Genome Sequencing Center (L. Hillier et al., in preparation), which has (as of 1 June 1997) contributed 179,134 sequences (97% of all mouse ESTs in dbEST). In both of these projects the IMAGE consortium [19] has been instrumental in collecting the cDNA libraries, arraying the clones, and making the clones available for sequencing and redistribution.

Table 1. Top 20 organisms represented in dbEST (numbers of EST sequence entries as of 1 June 1997)

Organism	ESTs
Homo sapiens	716,351
Mus musculus	183,857
Arabidopsis thaliana	31,174
Caenorhabditis elegans	30,196
Oryza sativa	12,899
Drosophila melanogaster	9,206
Brugia malayi	8,408
Toxoplasma gondii	8,318
Rattus sp.	4,822
Saccharomyces cerevisiae	3,042
Trypanosoma brucei rhodesiense	2,450
Caenorhabditis briggsae	2,424
Schistosoma mansoni	2,375
Plasmodium falciparum	2,011
Rattus norvegicus	1,828
Zea mays	1,757
Brassica napus	1,427
Sus scrofa	1,272
Onchocerca volvulus	1,207
Brassica campestris	965
Ricinus communis	750
Other species	7,342
Total	1,033,331

World Wide Web Resources

Database of Expressed Sequence Tags (dbEST):	http://www.ncbi.nlm.nih.gov/dbEST/
UniGene Human and Mouse collections:	http://www.ncbi.nlm.nih.gov/UniGene/
International RH Consortium Gene Map 1996:	http://www.ncbi.nlm.nih.gov/science96/
NCI Cancer Genome Anatomy Project (CGAP):	http://www.ncbi.nlm.nih.gov/ncicgap/

In order for EST sequencing to be maximally productive certain details of the library construction require some attention. For example, normalization procedures have been used to reduce the abundance of highly expressed genes so as to favor the sampling of rarer transcripts [20]. More recently subtraction techniques have been used to construct libraries depleted of clones already subjected to EST sampling [21]. But although these techniques make it more efficient to find transcripts which are at low abundance in a particular tissue, many genes will still be missed because they are simply not expressed at all in the particular tissue. Thus the key to obtaining sequence tags from the maximal number of genes is to sample from as many different tissue sources as possible. One noteworthy new project is the Cancer Genome Anatomy Project recently initiated by the United States National Cancer Institute [22]. One component of Cancer Genome Anatomy Project is the analysis of cDNAs from a wide variety of cancer tissues and their normal counterparts.

Cataloging the pieces

Using ESTs for the construction of a gene catalog poses a challenge. With over 700,000 human ESTs it is clear that many genes will be multiply represented. In one sense this can be considered useful because it could allow multiple sequencing reads to be assembled to produce longer and more accurate sequences. Unfortunately, it can be difficult to decide which sequences should be merged. When comparing the sequences, some amount of mismatching should be allowed due to the single-pass nature of the sequences. However, some members of gene families contain extended regions with very few differences. The possibility of alternative splicing patterns adds a new level of complexity to the problem. On top of this, some clones are derived from partially spliced pre-mRNAs, which are nearly indistinguishable from true splicing variants. Under these circumstances, each piece of the larger Puzzle of Life becomes a smaller puzzle in itself.

Despite the difficulties any level of organization and validation that can be

achieved for ESTs is clearly advantageous. One approach carried out at the Institute for Genome Research is to apply standard software designed for merging sequencing reads– essentially treating the problem as a large-scale fragment assembly project [14]. By this process ESTs have been merged to form 62,808 "tentative human consensus" (THC) sequences, with an additional 175,563 ESTs remaining as nonmatching (as of 21 April 1997). It should be noted that these numbers do not necessarily reflect the number of genes which have been sampled because there can be multiple nonoverlapping patches of sequence originating from a single mRNA. Nevertheless, THCs are useful for sequence database searching because much of the obvious redundancy is eliminated, which speeds up the search and reduces the volume of output.

Another approach to developing a gene catalog has been used in the construction of UniGene [23, 24]. In this case individual genes, as opposed to sequence contigs, provide the organizing framework. Using automated procedures, ESTs and full-length mRNAs from characterized genes are partitioned into sets, or "clusters," that are very likely to represent distinct genes. In addition to strong sequence similarity, clone identifiers can be used to group ESTs derived from the same cDNA even when their sequences do not overlap. The presence of an authentic 3' mRNA terminus can be used to "anchor" the sets, thereby avoiding multiple disjoint sets for the same gene. The number of anchored sets, currently 62,421 (as of 15 Feb. 1997), provides a rough estimate of the number of genes so far sampled. UniGene has been used to select reagents for large-scale gene mapping [24, 25] and gene expression [26] studies (see below). Moreover, the gene-oriented nature of UniGene makes it a useful scaffold on which to hang the results of these studies. A mouse version of UniGene has recently been constructed to accommodate the increasing numbers of mouse ESTs, and large-scale sequence comparisons have been used to cross-reference human and mouse sequence clusters. Homology maps between the human and mouse genomes have been constructed [27] and cross-referenced human and mouse sequences should be of utility in extending these studies.

A framework for the puzzle

Determining the map positions of each gene provides a new level of organization for the gene catalog, a framework for the Puzzle of Life, telling us where to put each piece, despite the fact that we cannot yet see how they fit with one another. Most modern mapping methodologies center around the use of sequence-tagged sites (STSs) as unique landmarks of the genome [28]. It has been demonstrated that single-pass sequences provide suitable templates for the design of gene-based STSs [29]. For several reasons these sequences should ideally originate from the 3' ends of the transcripts. Several mapping technologies make use of rodent cells as carriers for human genomic material. Consequently cross-reactivity between human STSs and the rodent background may be an issue. Sequences near the 3' ends of transcripts are very likely to fall within untranslated regions, and it has been found that such regions show less cross-species conservation than do coding sequences [30]. In addition, the paucity of introns near the 3' termini means that markers designed from cDNAs will be colinear with genomic DNA. Hence observed PCR product sizes should correspond well to predictions.

As it became clear that ESTs represent a huge untapped source of gene-based material for mapping studies, an international consortium was established to perform this work in a coordinated fashion. A nonredundant set of 3' end sequences was selected from UniGene and distributed to the participating laboratories, which included the Whitehead Institute Center for Genome Research, the Sanger Centre, the Stanford Human Genome Center, Genethon, and the Wellcome Trust Center for Human Genetics. STS markers were developed from these sequences and mapped using primarily radiation hybrid (RH) techniques. The first report from the RH consortium provided map positions over 16,000 individual genes, roughly one-fifth of the total number of human genes [24]. In addition, approximately 1000 genetic markers from the Genethon genetic map [31] were included in this analysis, both to serve as a mapping framework and to allow gene positions to be related to genetic linkage information.

The gene map provides an important new tool for disease gene hunters. In recent years a common strategy has been positional cloning, in which identification of the gene of interest is based primarily on map position [32]. This is a time-consuming process which often involves some localization of cDNAs in the region under scrutiny [33]. Now, as soon as disease susceptibility can be localized to an approximate interval (or several intervals), a simple database query can be used to instantly generate lists of candidate genes. Given the density of the current map, there is about one chance in five that the disease gene will correspond to an EST that has already been localized.

Adding color to the pieces

The Puzzle of Life is starting to take shape now that many of its pieces have been identified and positioned within the overall framework. Unfortunately, most of these pieces are blank. For the genes which have been identified only as ESTs virtually no functional information is available, no indication of what role they play in the larger picture of human biology. Determining the function of each gene is the next major challenge of an emerging field which has been called "functional genomics." Although there are many aspects to understanding gene function, one approach involves simultaneous determination of mRNA levels for large numbers of genes and correlation of this information with different biological contexts, such as specific disease states and experimental treatments.

The sequences of the ESTs themselves have been used as an indirect means of compiling gene expression profiles [12, 34, 35]. Following putative identification of the sequences the number of ESTs observed for each gene are tabulated to give a relative expression level. However, most libraries have not been constructed with this goal in mind. The normalization and subtraction techniques described above would (by design) alter transcript representations. Other factors, such as size of the clone insert, could introduce additional biases. Furthermore, a library would have to be fairly extensively sampled for EST profiling to be sensitive to rare transcripts – an expensive proposition. A recently described technology called serial analysis of gene expression (SAGE) partially overcomes these limitations by focusing on uniformly short sequence tags. SAGE has recently been used to rapidly survey gene expression patterns associated with pancreatic, colon, and lung cancers [36]. Techniques based on direct sequencing of transcript-derived material have the advantage of not requiring the prior existence of a gene catalog.

With a catalog of cDNA sequences and clones in hand, a number of emerging technologies make it possible to simultaneously monitor the expression of tens of thousands of individual genes. Cloned cDNAs can be spotted in high-density microarray format on glass slides and used as hybridization targets for probes made by fluorescent labeling of mRNA samples [37]. This methodology has recently been used to identify genes of potential relevance for malignant melanoma [26] and rheumatoid arthritis [38]. An alternate approach, which has been commercialized by the company Affymetrix, makes use of a high-tech combination of oligonucleotide chemistry and photolithography (of the sort used in the fabrication of computer microcircuitry) to construct high-density arrays of oligonucleotides on a silicon wafer, which, again, can be used as hybridization targets [39]. Both of these methods generate huge amounts of data, which must then be linked back to the gene catalog so that the results can be interpreted.

Regardless of the assay system employed the accuracy of the results and the ability to compare expression profiles from one tissue sample to another requires careful attention to the preparation of tissues and mRNA samples. For example, gene products normally associated with blood are seen in many cDNA libraries for the simple reason that no attempt was made to remove blood from the tissue samples. It is fully expected that expression patterns in most tissues will display heterogeneity at the microscopic level, although this has traditionally been difficult to study. However, recently developed laser capture microdissection technology makes it possible to isolate small groups of well-defined cells for expression analysis and even cDNA library construction [40]. This will be a key technology for CGAP, where the aim is to understand the molecular changes that accompany the development of cancer.

Completing the Puzzle

A catalog of human gene sequences is a critical resource for both computational and experimental approaches to genome analysis, and it provides a framework for organizing and understanding the information collected from these studies. Continued sampling of cDNAs will allow for identification of the remaining genes and confirmation of those which have so far been seen only once. However, there is little doubt that the biggest impact will come from identifying a nonredundant set of full-length cDNAs and obtaining complete, high-quality sequences from them. This will eliminate current uncertainties in the gene sequence collection, provide protein sequences for functional and evolutionary studies, allow the unambiguous elucidation of transcription units in genomic sequences, and identify a reference set of cDNAs for use in experimental studies.

The solution to the puzzle may be several years away, but each step toward that goal is a new advance in our understanding of biology. Traditional studies in molecular biology have been somewhat reductionist, focusing on individual genes – single pieces of the larger picture. As localized regions of the puzzle come together, the interactions among genes will define pathways. Pathways will fit together into systems. Finally, the interactions of systems will fall into place to reveal the "big picture" of human biology.

Acknowledgements Special thanks to Ken Katz for helpful suggestions and critical review of the manuscript.

References

1. Fields C, Adams MD, White O, Venter JC (1994) How many genes in the human genome? Nat Genet 7:345–346
2. Antiquera F, Bird A (1993) Number of CpG islands and genes in the human and mouse genomes. Proc Natl Acad Sci USA 90:11995–11999
3. Koonin EV (1997) Big time for small genomes. Genome Res 7:418–421
4. Collins F, Galas D (1993) A new five-year plan for the U.S. human genome project. Science 262:43–46
5. Brenner S (1990) The human genome: the nature of the enterprise. Ciba Found Symp 149:6–12
6. Adams MD et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656
7. Sikela JM, Auffray C (1993) Finding new genes faster than ever. Nat Genet 3:189–191
8. Boguski MS, Tolstoshev CM, Bassett DE Jr (1994) Gene discovery in dbEST. Science 265:1993–1994
9. Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST – database for "expressed sequence tags." Nature Genet 4:332–333
10. Benson DA, Boguski M, Lipman DJ, Ostell J (1996) GenBank. Nucleic Acids Res 24:1–5
11. Altschul SF et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
12. Matsubara K, Okubo K (1993) cDNA analyses in the human genome project. Gene 135:265–274
13. Houlgatte R et al (1995) The Genexpress Index: A resource for gene discovery and the genic map of the human genome. Genome Res 5:272–304

14. Adams MD et al (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. Nature 377:3–17
15. Hillier L et al (1996) Generation and analysis of 280,000 expressed sequence tags. Genome Res 6:807–828
16. Waterston RW et al (1992) A survey of expressed genes in *Caenorhabditis* elegans.
17. McCombie WR et al (1992) *Caenorhabditis* elegans expressed sequence tags identify gene families and potential disease gene homologues. Nat Genet 1:124–131
18. Newman T et al (1994) Genes galore: A summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. Plant Physiol 106:1241–1255
19. Lennon G, Auffray C, Polymeropolous M, Soares MB (1996) The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. Genomics 33:151–152
20. Soares MB et al (1994) Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci USA 91:9228–9232
21. Bonaldo M, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791–806
22. Strausberg RL, Dahl CA, Klausner RD (1997) New opportunities for uncovering the molecular basis of cancer. Nat Genet [Suppl]:415–416
23. Boguski MS, Schuler GD (1995) ESTablishing a human transcript map. Nat Genet 10:369–371
24. Schuler GD et al (1996) A gene map of the human genome. Science 274:540–546
25. Hudson TJ et al (1995) An STS-based map of the human genome. Science 270:1945–1954
26. DeRisi J et al (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. Nat Genet 14:457–460
27. DeBry RW, Seldin MF (1996) Human/-mouse homology relationships. Genomics 33:337–351
28. Olson M, Hood L, Cantor C, Botstein D (1989) A common language for physical mapping of the human genome. Science 245:1434–1435
29. Wilcox AS, Khan AS, Hopkins JA, Sikela JM (1991) Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. Nucleic Acids Res 19:1837–1843
30. Makalowski W, Zhang Z, Boguski MS (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. Genome Res 6:846–857
31. Dib C et al (1996) A compreshensive genetic map of the human genome based on 5,254 microsatellites. Nature 380:152–154
32. Collins FS (1995) Positional cloning moves from perditional to traditional. Nat Genet 9:347–350
33. Brennan MB, Hochgeschwender U (1995) Commentary: So many needles, so much hay. Hum Mol Gen 4:153–156
34. Adams MD, Kerlavage AR, Fields C, Venter JC (1993) 3,400 new expressed sequence tags identify diversity of transcripts in human brain. Nat Genet 4:256–267
35. Gress TM et al (1996) A pancreatic cancer-specific expression profile. Oncogene 13:1819–1830
36. Zhang L et al (1997) Gene expression profiles in normal and cancer cells. Science 276:1268–1272
37. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with complementary DNA microarray. Science 270:467–470
38. Heller RA et al (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 94:2150–2155
39. Lockhart DJ et al (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotech 14:1675–1680
40. Emmert-Buck MR et al (1996) Laser capture microdissection. Science 274:998–1001